DOCUMENT RESUME

ED 221 570 TM 820 594

AUTHOR Legg, Sue M.
TITLE The Use of Precalibrated Item Bank to Establish and Maintain Cutoff Scores: A Case Study of the Florida Teacher Certification Examination.
PUB DATE Mar 82
NOTE 14p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; Difficulty Level; Field Tests; Item Analysis; *Item Banks; *Latent Trait Theory; Scaling; *Scoring Formulas; *Statistical Analysis; Teacher Certification; Test Bias; *Test Construction; Testing Problems; Test Items
IDENTIFIERS *Calibration; *Florida Teacher Certification Examination; Logit Analysis; Rasch Model; Rasch Scaled Scores

ABSTRACT

A case study of the Florida Teacher Certification Examination (FTCE) program was described to assist others launching the development of large scale item banks. FTCE has four subtests: Mathematics, Reading, Writing, and Professional Education. Rasch calibrated item banks have been developed for all subtests except Writing. The methods used to evaluate the stability of the score scale as it relates to a cut score based upon a logit value are explained. Cut off levels were established in terms of a logit ability scale based on field test data. Linking items were chosen so that difficulty levels were centered at the cut score. Item selection was based upon closeness of fit to the Rasch Model and item difficulty. The linking items for the Professional Education subtest were of primary concern. Results showed no bias in the test due to differences in curriculum. (DWH)

The Use of Precalibrated Item Bank to Establish and Maintain Cutoff Scores: A Case Study of the Florida Teacher Certification Examination

Dr. Sue M. Legg

Office of Instructional Resources
University of Florida
Gainesville, Florida

Paper presented at the annual meeting of the National Council of Measurement in Education

March 1982

The Use of Precalibrated Item Bank to Establish and Maintain Cutoff Scores: A Case Study of the Florida Teacher Certification Examination

Applications of Rasch analysis to test development are attracting considerable interest. Since relatively few measurement specialists have experience in the actual development and maintenance of precalibrated item banks, the case study of the Florida Teacher Certification Examination program may be helpful to others launching large scale item banks. The measurement problems of field test design, item independence, scale score stability and setting cutting scores are all confronted in this case study.

Description of the Item Bank

The Florida Teacher Certification Examination (FTCE) has four subtests: Mathematics, Reading, Writing and Professional Education. Rasch calibrated item banks have been developed for all but the Writing subtest. The original item banks were calibrated using a field test design in which seven test forms were administered. Each form included items grouped into the three subtests. Three sets of linking items were woven into each form. Test forms were equated in an anchor design by adding the linking constant to the unadjusted item logit difficulties.

Experimental items have been added to the bank on two occasions after field testing in the regular April and August administrations. The practical problems of cost efficiency and adequacy of measurement were addressed by using a common set of scored items across

forms with different sets of experimental items in each form. Each form was calibrated separately, and the linking constant was derived by averaging the difference in difficulty estimates of the scored samples. The adjusted values of the item logits for the scored items became the base for linking the experimental items. This procedure was a variation of the method described by Ryan in Equating New Test Forms to an Existing Test. By creating an average logit value for each item across forms and then computing the average difference between this set of items and their base value in the item bank, an adjustment factor is established to link the experimental items in each form to the item bank.

A computer based history bank has been developed to trace the stability of the item logits across administrations. The bank includes the field test logit as well as the unadjusted and adjusted logits from each administrations. Fit statistics are also included for each item. The bank is used to update another item file which contains the content codes and other attributes associated with each item. The attribute bank is organized by the content codes to allow the test developer to monitor the number and difficulty of items within each content area.

The techniques and problems described in this paper are familiar to those measurement specialists charged with the development and maintenance of large scale precalibrated item

banks. Many of these procedures have been presented in papers by Ryan, Mead, Ingebo and others. The contribution of this paper is the extension of these methods to evaluate tne stability of the score scale particularly as it relates to a cut score based upon a logit value.

Setting Cut Off Levels

The issue of the stability of the scale is related directly to the fact that all passing scores were set in terms of the logit ability scale based on the field test data. Using the method described in a companion paper, passing scores were set at 1.4 logits for reading, 1.0 logits for mathematics and .25 logit for professional education. Since difficulty levels increase with the value of the logit ability scale, the reading items require the highest level of proficiency and the professional education items have the lowest level. Normally, the higher the passing score, the greater the number of items that a candidate must get correct. In a calibrated item bank, the score standard need not be tied to a specific number of correct responses. Raw score equivalence to a cutoff may vary as a function of the difficulty of the items when the cutoff is based upon an ability logit value.

Linking Design

Linking items were selected based upon the closeness of fit to the Rasch Model and the item difficulty. In order to

enhance the reliability of the cutoff, linking items were chosen so that the difficulties were centered at the cut score. The number of linking items was assigned to be proportional to the number of items in each content area within subtests. This design necessitated that the total number of linking items was large. Approximately twenty percent of the items were designated as links.

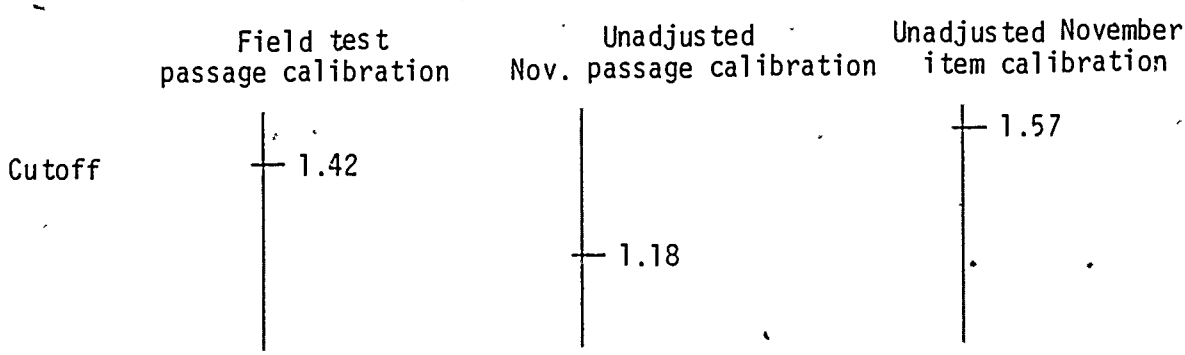
Establishing the Scale

A conservative approach to test development was taken for the Teacher Certification Examination. Attempts were made to stabilize the difficulty of each examination by selecting items which have an average logit value of zero. In this way, each test administration should have forms that are approximately equal in difficulty. Small variations in the difficulty of individual examinations are corrected through the linking procedures to the base scale derived from the field test. However, the flexibility in adjusting the number of items in a given subtest or allowing the raw score equivalence to the cutoff vary with the difficulty of the items is a decided advantage particularly when the item banks are being built and the number of items in given content areas may be small.

The problem of reducing the number of items within subtests did occur after the November 1981 administration. It was apparent that the testing time could not be extended in

order to include experimental items and a separate administration of experimental items was too cumbersome and costly. The decision was made to reduce the Professional Education subtest from one hundred items to eighty items. The Reading subtest was reduced by one passage and ten items. This change was made without altering the score scale. The ability scale derived from eighty or one hundred items was the same.

Another scaling problem occurred when the decision was made to alter the method of calibration for the Cloze reading test. This subtest was initially calibrated by passage, i.e. one passage score was obtained for the set of ten items associated with each passage. It was later decided to recalibrate on the item level which necessitated an adjustment to correct the scale for the method of calibration. The correction factor for the recalibration was calculated by finding the difference in the unadjusted ability logits at the cut score for the November calibrations by passage and by item. This difference was added to the adjustment of the November passage scale to the field test scale as shown below. This additional step was necessary because the original field test data was not available for recalibration.



In order to bring the November item calibration scale to the passage scale, a factor of -.390 was required (1.57 - 1.18). The adjustment of the November passage scale to the field test was .243 (1.42 - 1.18). The manipulation of the scales was completed by combining the two factors of -.390 and .243 to create the final adjustment of -.147.

Reliability of the Cutoff

Since the passing score determined eligibility for teacher certification, it was essential that the examination was reliable at the cutoff as well as internally consistent. Reliability was assessed at the cutoff by computing the standard error of each subtest at the passing score and by estimating the reliability of the cut scores using the Brennan Kane index. These estimates ranged between .92 and .96 for each subtest. Since the index is sensitive to the average difficulty of the items, reliability of the cutoff increases as the average difficulty decreases.

Maintaining the Score Scale

The steps outlined by Ron Mead in his paper entitled Basic Ideas in Item Banking have been followed to investigate the within form fit, the within link fit and the between link fit. These analyses were conducted to monitor the stability of the item calibrations and the score scale. Specific questions were raised about the stability of the items during the

development of the item bank and the early administrations of the examination which guided the analysis. These questions concerned the viability of the field test results and the possible effects on the calibrations of different candidate populations.

Any field test analysis has the possibility that the candidates writing the examination have not taken the items seriously. The item calibrations may yield difficulty estimates that are too high simply because of a cavalier attitude by some students. When cutoff scores are based upon field test data, the possibility exists that they may be too low in spite of the attempt to set the standards on the basis of the skills rather than on the pass rate. No standard can be set without the consideration of its possible effect. Therefore, the standards set for the Certification examination have been monitored to determine whether or not the ability reflected in the field test was representative of the ability scores derived from each subsequent administration. To date it appears that there is a slightly higher percentage of candidates than might have been expected that exceeds the cutoff based on field test results.

The stability of the item calibrations was also questioned due to possible differences in the candidates' knowledge of the content. Even though the Rasch model is sample free, it is not

impervious to the effect of differences in curriculum. The specifications for the examination were based upon the curriculum in teacher education programs within the state. However, substantial numbers of candidates from other states come to Florida to teach and must sit the examination. The reading and mathematics skills could be considered to be basic and therefore common to all students. The skills in professional education may vary in emphasis depending upon the curriculum of an institution. Even within the state there was considerable debate about the validity of the examination for vocational teachers. It was possible that the calibrations could change significantly during early administrations when the proportion of out of state and vocational educational candidates was large.

In order to investigate this possibility, linking item difficulties were plotted across administrations to ensure that they were parallel. Comparisons of adjusted item logit values were made, and substantial changes in item calibrations were examined. Standard item analyses were conducted for several categories of examinees to evaluate the possibility of differential success rates on items that could be tied to differences in curriculum.

Monitoring the Linking Items

While all linking items were monitored in each subtest, the links for the Professional Education subtest were of primary concern. The possible effect of differences in the educational

training of the candidates would most likely be shown in an analysis of this subtest. These specific questions guided the inquiry:

1. Were any shifts in values due to ambiguity in the items?
2. Was there a systematic shift in difficulty from the field test?
3. Was there a pattern of shifts in logit values that could be related to curriculum differences?

The Professional Education subtest contains twenty eight linking items distributed proportionately across twenty one competencies within six content areas:

Instructional Objectives

Evaluating, Recording, and Reporting Student Progress

Classroom Management

Learning and Teaching

Development of Students

Instructional Materials

A difference of more than one half logit from the field test was set as a significant change in value. The plot (See Figure 1) shows that fourteen items had changed in value. Eight items increased in difficulty and six decreased. However, five of these items were skewed in only one of their four administrations. Four of these five skewed logits appeared in either the field test or the first administration. Outliers from these administrations were expected and were relatively few in number.

There appeared to be no relationship between the content and the shift in logit values. The fourteen items represented ten different competencies within all six content areas. The fluctuation of the logits was random within the curriculum. Since similar results were found in the Mathematics and Reading subtests, the likelihood of any pattern being found was small.

The item analyses revealed that out of state candidates generally performed better than in-state students on all subtests. The changes in the values of the linking items were unrelated to the proportion of out of state students in each administration.

Candidates for vocational-technical certification did less well on all subtests than in-state or out of state groups. However, they performed better on professional education items than on basic skills. Thus, their results were most likely due to a generally lower level of academic skills.

All of the evidence points to the conclusion that there is no bias in the test due to differences in curriculum. Each content area has similar success rates for each of the population categories discussed.[1] The proportion of items necessary to pass the examination remains relatively constant. It varies only with small differences in test difficulty. Difficulty is not related to differences in content or curriculum.

[1] Figure 2, page 12.

Figure 1. Plot of Logits: Professional Education

1 = Field
2 = November
3 = April
4 = August

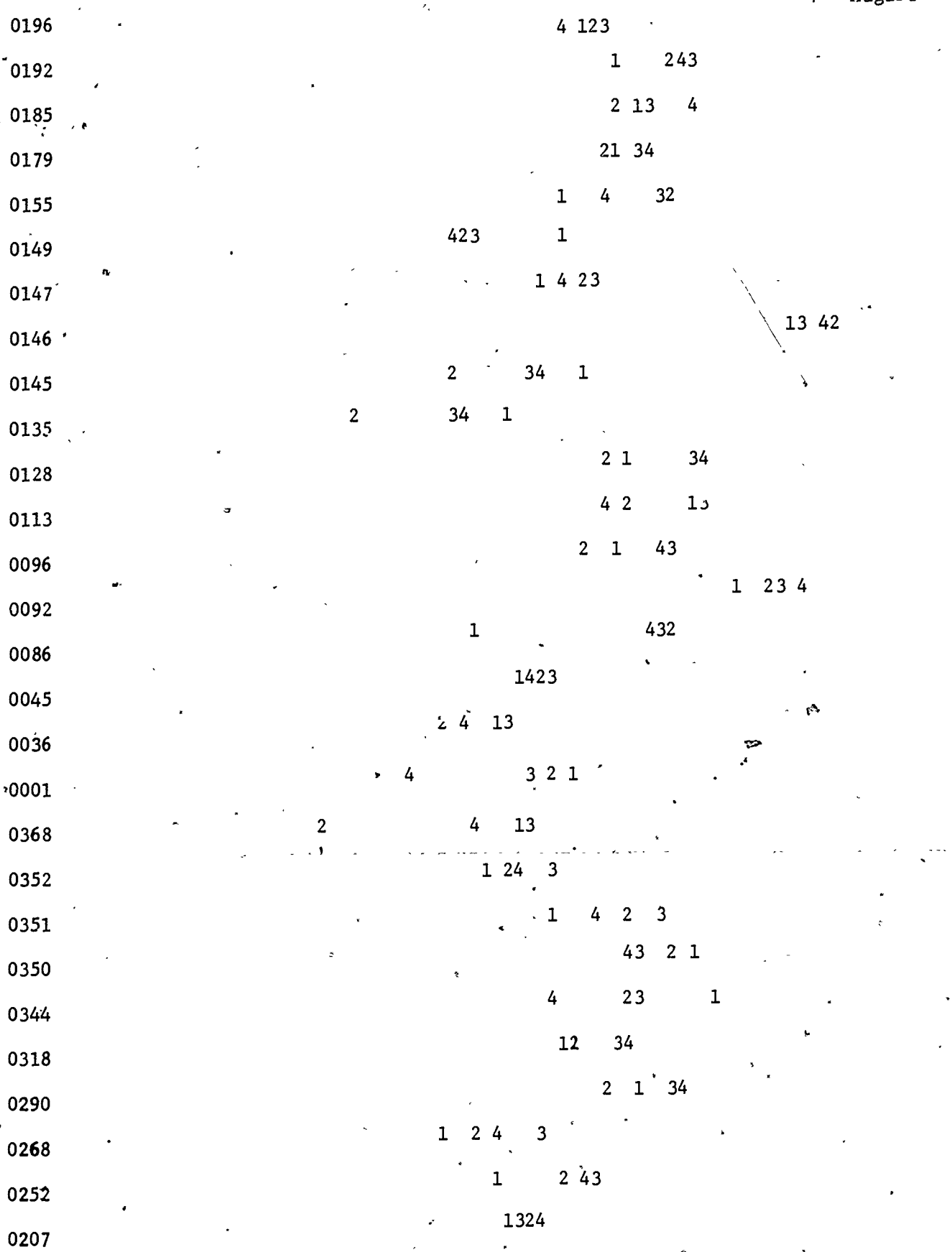


Figure 2. Percent of Items Correct by Content Base Categories

	(1) Management	(2) Development	(3) Measurement	(4) Materials	(5) Objectives	(6) Learning
November	73	75	73	77	66	73
April	76	73	80	77	80	77